Categorical Term Descriptor: A Proposed Term Weighting Scheme for Feature Selection

Bong Chih How[1], Narayanan Kulathuramaiyer[1], Wong Ting Kiong[2]
Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia
[1]{chbong,nara}@fit.unimas.my
[2]kiong_ge_wtk@yahoo.com

Abstract

This paper proposes a term weighting scheme, Categorical Term Descriptor (CTD), for feature selection in automated text categorization. CTD is an adaptation of the Term Frequency Inverse Document Frequency (TFIDF). We compared the performance of the proposed method against classical methods such as Correlation Coefficient, Chi-Square and Information Gain using the Multinomial Naïve Bayes and the Support Vector Machine (SVM) classifiers on the Reuters(10) and Reuters(115) variants of Reuters-21578 dataset. Despite its simplicity, CTD has proven to be promising for both local and global feature selection. CTD works best for the Reuter(10) as a stable local FS method

1. Introduction

Term weighting schemes (TWS) such as Term Frequency Inverse Document Frequency (TFIDF) has been widely used to identify significant terms to describe documents [6][14]. Determining the effectiveness of particular term weighting scheme remains an open research question, as there are many variations of such schemes that have been proposed [3][4]. In this paper, we propose an adapted term weighting scheme, Categorical Term Descriptor (CTD), to be used as a feature selection (FS) method. In order to validate its performance, we compare CTD's ability in performing local and global feature selection with well-known methods such as Correlation Coefficient, Chi-Square and Information Gain FS [1][2]. The quality of features selected is closely dependent on the characteristics of datasets used. Two variants of the Reuters are thus required. Experiments have been conducted using both the Multinomial Naïve Bayes (MNB) and Support Vector Machine (SVM), to determine the actual effects of CTD. Another important consideration is to produce a computationally economical FS method.

2. TWS for Feature Selection

The success of text categorization highly depends on the efficiency and accuracy of feature selection [1][2]. Feature selection (FS) is generally used to perform dimensionality reduction. It employs the "importance" of terms in order to identify a subset of the original set of terms that effectively characterize the category. Each term is weighted and scored based on the overall corpus statistics. In contrast, term weighting schemes are typically used to determine significant keywords to be used in document indexing, rather than to capture descriptive terms to describe category. Using the bag-of- words approach, terms derived from text documents can be directly identified as features.

Intuitively, FS can be performed in two distinct ways. Local selection chooses features from each category of documents separately. Global FS selects features across all categories. A globalization technique specifically assigns a global score to each respective category via numerical functions such as sum, average or maximum. Global FS techniques do not take into consideration application specific data characteristics in extracting features across all the categories.

3. Feature Selection

3.1. Information Gain (IG)

IG is considered to be the best known measure of feature quality in text categorization [1][2][10]. We adopt the definition of IG based on [1], it defines

$$IG(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) - p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) - p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

3.2. Chi-Square (CHI)

CHI is another widely used statistical measurement used FS. It measures the degree of independence between a term t and a category c. The higher the CHI

Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)
0-7695-2415-X/05 $20.00 © 2005 IEEE